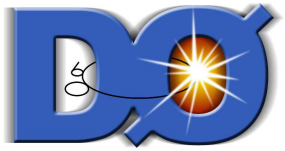
Grid Experiences in DØ



Daniel Wicke
(Bergische Universität Wuppertal)

Outline

- Prolog: DØ data handling
- Data (re)processing on the Grid
- GIF: Transfer to LCG

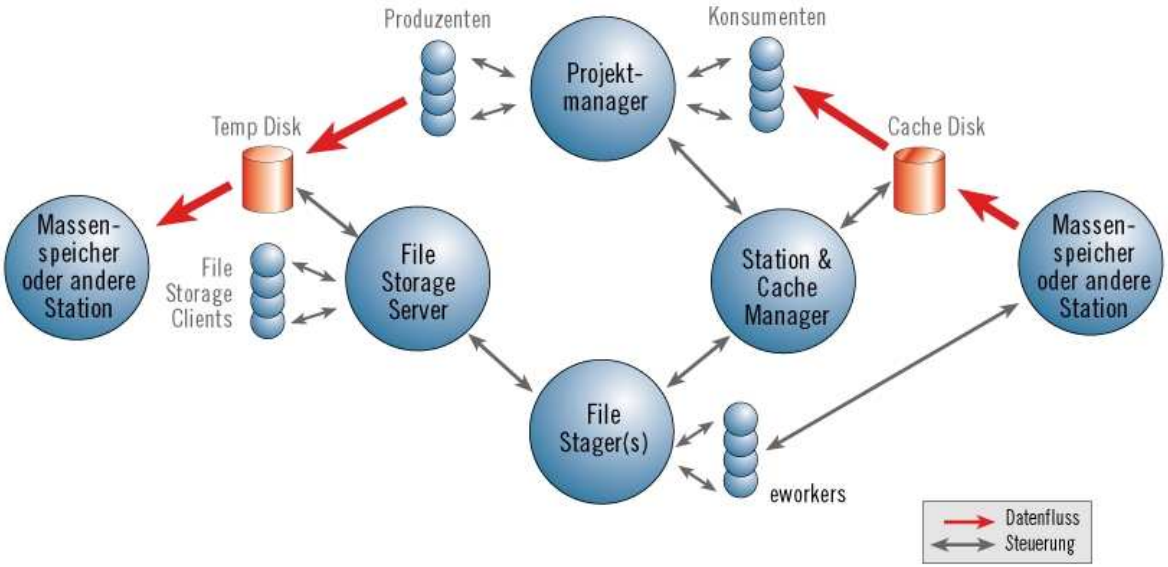
Prolog: DØ Data Handling

Sequential Access through Metadata: SAM

The order of events in the dataset has no meaning.

Optimisation

- Don't loop through file lists.
- Request datasets.
- The order in which files corresponding to a dataset are processed may change.
- The system optimises the order to minimise tape access and tape mounts.



WAN Transport

`sam_cp` as generic interface to `GridFTP`, `bbftp`, tape access . . .

The Computing Task

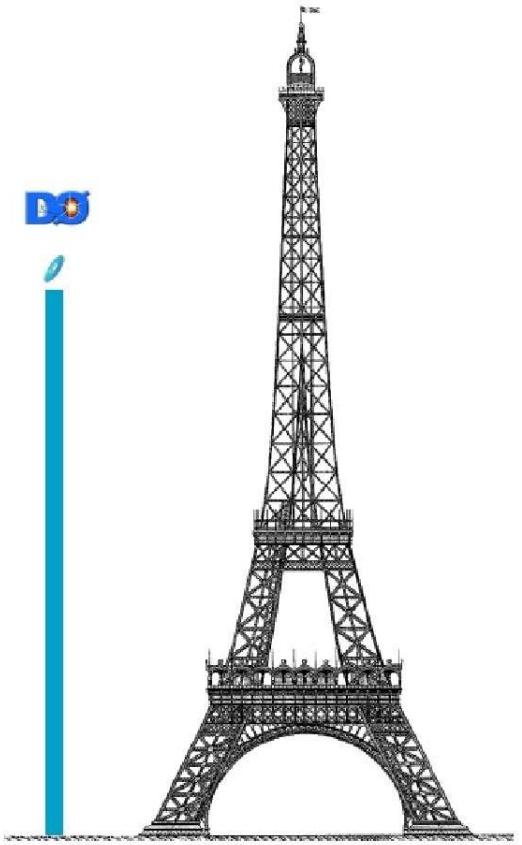
	p17 reprocessing	p14 reprocessing
Luminosity	$470\,\mathrm{pb}^{-1}$	$100\,\mathrm{pb}^{-1}$
Events	1G	300M
Rawdata $250\mathrm{kB}$/Event	250TB	75TB
DSTs $150\mathrm{kB}$/Event	150TB	45TB
TMBs $70\mathrm{kB}$/Event	70TB	6TB
Time $50s$/Event	$20,000$months	6000months
(on 1GHz Pentium III)	3400CPUs for 6mths	2000CPUs for 3mths
Remote processing	100%	30%

Central Farm (1000CPUs) used to capacity with data taking.

The Computing Task

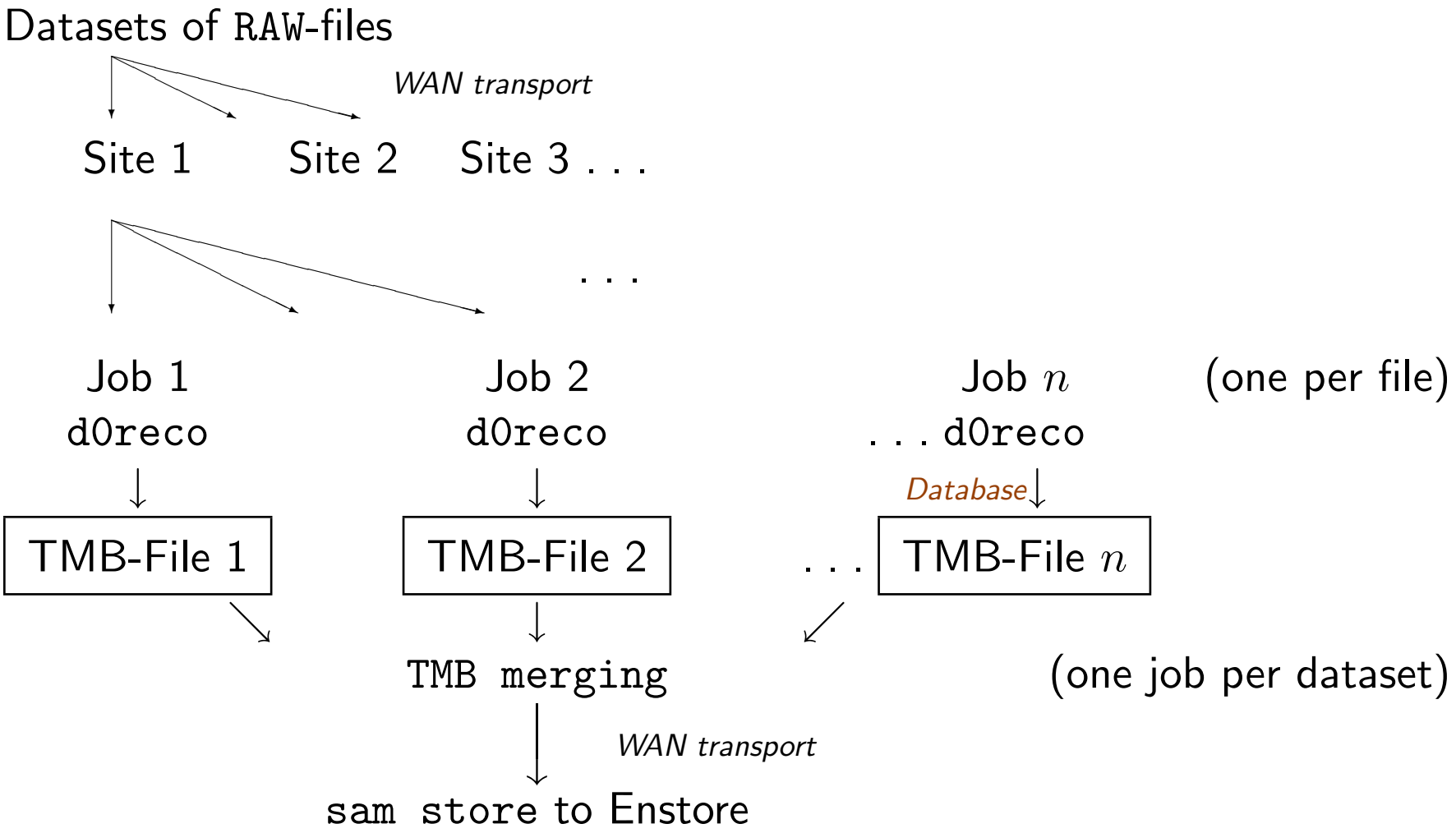
	p17 reprocessing
Luminosity	$470\,\mathrm{pb}^{-1}$
Events	1G
Rawdata $250\mathrm{kB}$/Event	250TB
DSTs $150\mathrm{kB}$/Event	150TB
TMBs $70\mathrm{kB}$/Event	70TB
Time $50s$/Event	$20,000$months
(on 1GHz Pentium III)	3400CPUs for 6mths
Remote processing	100%

A stack of CDs as high as the Eiffeltower



Application flow

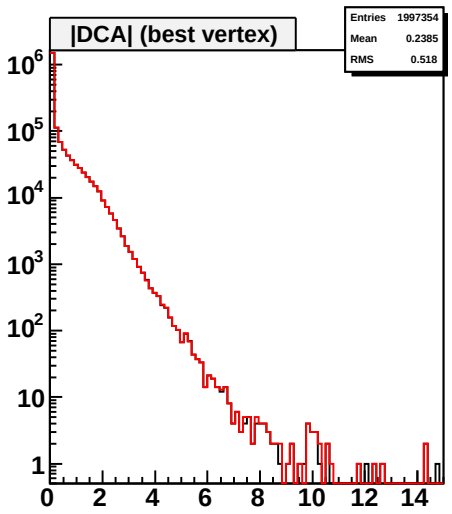
Overview



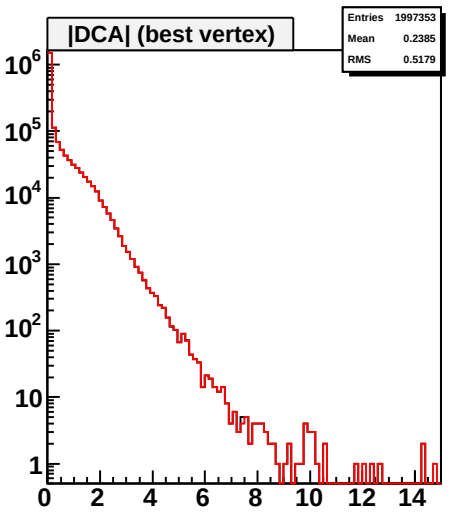
Certification of Sites and Code

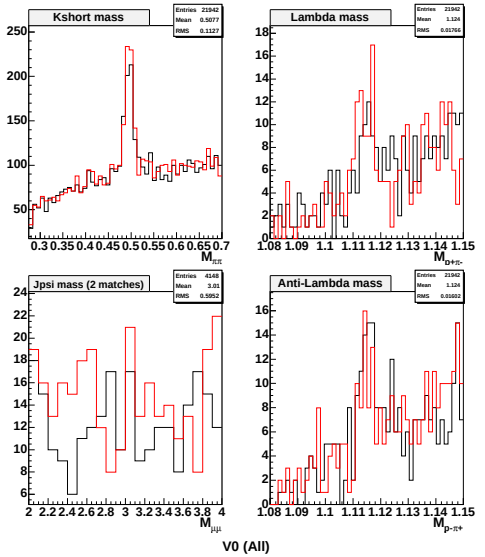
- Compared SAMGrid production to conventional production on `d0farm`.
- Compared SAMGrid production at each site to `d0farm` production.
- Compared merged to unmerged TMBs at each site.

Lead to significant improvements in `recocert`



D0Farm JIM vs Lyon JIM





D0Farm Std vs JIM

Error Handling and Recovery

Beside unrecoverable crashes of `d0reco` there will be *random* crashes.
(Network outages, file delivery failures, batch system crashes/hangups, worker-node crashes, filesystem corruption...)

Book-keeping

1. of succeeded jobs/files
   needed to assure completion without duplicated events.

   SAM is used avoid data duplication and to define recovery jobs.

2. of failed jobs/files
   needed to trace problems in order fix bugs and to assure efficiency.

   JIMs XML-DB is used to ease bug tracing and provide fast recovery.

The steering software: D0Repro

- Create neccessary Gridjob with minimal manual parameters
- Automatic prevention of duplicate production
- Evaluation of database for status display

The Autopilot

- Automatic execution of required command chain
- Automatic error recovery (up 5% error rate)
- Allows to run production several days automatically.

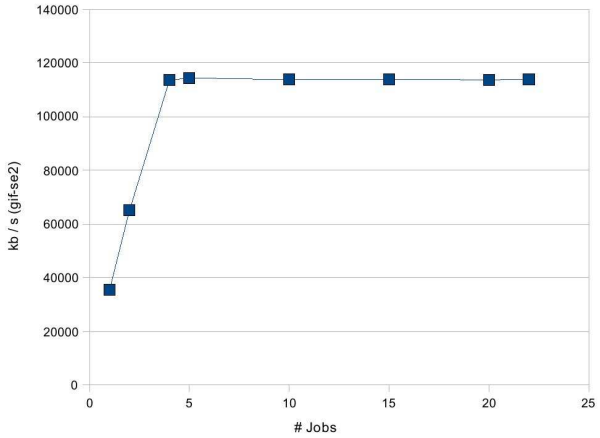
Follow up campaigns

- another reprocessing run by students
- standard for primary processing of data

GIF: Transfer DØ experiences to LCG

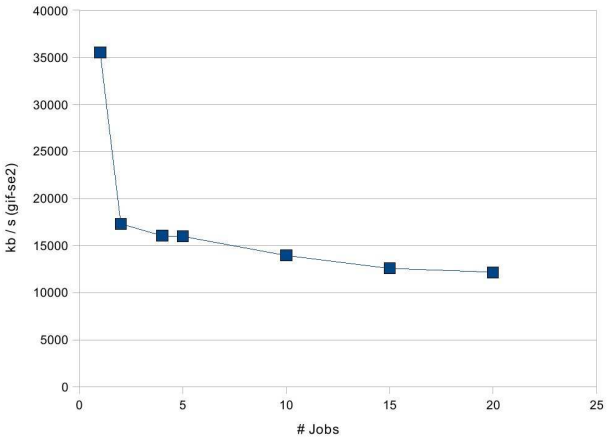
Experiences in DØ

- File access of many parallel batch-jobs problematic
- Single spikes may overload the system
- DØ solution: serialisation (fcp)



Transfer to LCG

- Assembled a Test-Cluster
  - Virtual nodes (XEN)
- Verified the preformence of the system
- Implemented serialising filesystem



WAN Tests here now.

Conclusions

- DØ has aimed for distributed/grid computing since the beginning of RunII
  - Production of MC simulation was only done remotely
  - Data production was added later,
  - Over time more and more gridified.

- DØ was the first HEP experiment to process data on non-dedicated sites.
  - It was the largest HEP grid project with data at the time.

- DØ grid is based on Globus (as LCG), but
  - has a centralised permanent storage
  - uses data transport to the job
  - creates multiple batch-job per grid job

- The data access problem is still applicable